Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Transfer learning based clinical concept extraction on data from multiple sources

Xinbo Lv, Yi Guan *, Benyang Deng

School of Computer Science and Technology, Harbin Institution of Technology, Harbin, Heilongjiang 150001, China

ARTICLE INFO

Article history:
Received 6 July 2013
Accepted 6 May 2014
Available online xxxx

Keywords:
Clinical concept extraction
Transfer learning
TrAdaBoost
Bagging
Machine learning

ABSTRACT

Machine learning methods usually assume that training data and test data are drawn from the same distribution. However, this assumption often cannot be satisfied in the task of clinical concept extraction. The main aim of this paper was to use training data from one institution to build a concept extraction model for data from another institution with a different distribution. An instance-based transfer learning method, TrAdaBoost, was applied in this work. To prevent the occurrence of a negative transfer phenomenon with TrAdaBoost, we integrated it with Bagging, which provides a "softer" weights update mechanism with only a tiny amount of training data from the target domain. Two data sets named BETH and PARTNERS from the 2010 i2b2/VA challenge as well as BETHBIO, a data set we constructed ourselves, were employed to show the effectiveness of our work's transfer ability. Our method outperforms the baseline model by 2.3% and 4.4% when the baseline model is trained by training data that are combined from the source domain and the target domain in two experiments of BETH vs. PARTNERS and BETHBIO vs. PARTNERS, respectively. Additionally, confidence intervals for the performance metrics suggest that our method's results have statistical significance. Moreover, we explore the applicability of our method for further experiments. With our method, only a tiny amount of labeled data from the target domain is required to build a concept extraction model that produces better performance.

1. Introduction

Clinical documents are valuable resources in which abundant personalized health information, such as symptoms, medicines and tests, is recorded by physicians in natural language. As a subtask of automatic acquisition of knowledge from these unstructured clinical texts, concept extraction aims to identify words and phrases that stand for clinical concepts from the narrative texts in clinical documents. This is the key component of text processing systems for understanding the content of clinical documents. Only when clinical concepts are correctly identified can other more complex tasks, such as concept relation extraction, assertion classification, co-reference, health information retrieval and health information recommendation, be performed effectively.

In the biomedical literature domain, research similar to concept extraction has been conducted in named entity recognition tasks such as gene name recognition [1]. However, research on clinical concept extraction for clinical documents appears to be rather sparse. One important reason for the lag of clinical concept extraction is the lack of shared annotated clinical documents due to patient privacy and confidentiality requirements. Fortunately, efforts to construct de-identified clinical documents are finally allowing studies on clinical concept extraction. For example, the 2010 Informatics for Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) challenge [2] provided a total of 394 training documents, 477 test documents, and 877 un-annotated documents for all three tasks. However, annotated clinical documents are always scarce and are created by a number of different institutions; the 2010 i2b2/VA challenge's data consist of four sets from three institutions. Such small-scale data sets limit the performance of a statistical machine learning model. One solution to this problem is to increase the training data sets by gathering data from multiple sources. Nevertheless, different vocabularies and writing styles of multiple sources make the combined data sets heterogeneous, to which the statistical machine learning model is sensitive. Specifically, the marginal probability distributions of words in clinical texts from different institutions are not equal, which violates the traditional machine learning's basic assumption: the training and test data should be under the same distribution. Therefore, a learner trained by one institution's data may perform worse when it is applied to data from another institution. The normal way of tackling this problem is to annotate data from the new institution,

* Corresponding author. Address: Mailbox 321, West Da-zhi Street 92, Harbin, Heilongjiang 150001, China.
E-mail address: guanyi@hit.edu.cn (Y. Guan).

http://dx.doi.org/10.1016/j.jbi.2014.05.006
1532-0464/

but this is always expensive and time-consuming. Abandoning old data would also be a waste.

The objective of this paper is to discuss approaches and strategies for clinical concept extraction from multiple sources. Using a training data set with a different distribution from one institution, we build a clinical concept extraction model for data from another institution. Transfer learning is a family of algorithms that can relax the traditional machine learning's same-distribution assumption. It leverages and transfers knowledge from the source domain to the target domain, and in this way, helps improve the model when the target domain's training data are insufficient. Specifically, we apply an instance-based transfer learning method – TrAdaBoost [3] – to the clinical concept extraction task. TrAdaBoost aims to re-weight the instances in the source domain in order to decrease the diversity between the data of the source domain and the target domain. It was originally created to solve binary classification problems, and we apply it to the sequence labeling problem with multiple labels. Additionally, to avoid the negative transfer problem caused by the over-discarded risk of TrAdaBoost, we integrate Bagging with TrAdaBoost to provide a "softer" weight update mechanism. Two data sets, BETH and PARTNERS, from the 2010 i2b2/VA challenge, as well as one data set we built by combining BETH and a biomedical literature data set (BIOLITERATURE), are used to verify the effectiveness of our method's transfer ability. Experiments show that with only a small amount of annotated training data from the target domain, our framework outperforms the baseline method, which simply combines data from the source domain and data from the target domain as training data.

2. Background

Methods for clinical concept extraction generally fall into three categories: dictionary-based methods, rule-based methods and statistical machine learning methods [4].

Dictionary-based methods search through dictionaries such as UMLS [5] and SNOMED-CT [6] to extract clinical concepts. MedLEE [7] is a typical system that uses a domain-specific vocabulary and semantic grammar to extract and encode clinical information in narrative reports. A structured representation is then constructed by these clinical terms. It is adapted to extract the concepts in clinical documents, and these concepts are mapped to semantic categories and semantic structures [8]. MetaMap [9] is also an early dictionary-based program developed at the National Library of Medicine (NLM) to recognize and categorize entities in texts from the biomedical domain and then to map them to UMLS Metathesaurus. It is applied to both IR and data mining applications; additionally, it is used to index the biomedical literature at the NLM. Systems described in [10–12] also adopt dictionary-based methods. The advantage of these methods is that they are easy to implement, while the disadvantage is that they suffer from low recall since many concepts may fail to be covered by the dictionary.

Rule-based methods require experts to define hand-coded rules or regular expressions for the extraction task. For example, in the sentence "systemic granulomatous diseases such as Crohn's disease or saroiaosis", the phrase "such as" implies that "Crohn's disease" and "saroiaosis" are disease names. Long [13] used regular expression to extract the diagnoses and procedures from the past medical history and discharge diagnoses in the discharge summary. Turchin et al. [14] designed a software tool to extract blood pressure values and anti-hypertensive treatment intensification from the texts of physician notes; regular expressions are also employed in their work. Rule-based methods are always difficult to achieve and time-consuming because rules have to be collected by hand.

In recent years, more and more researchers have resorted to statistical machine learning methods for clinical concept extraction. Several models, such as the Hidden Markov Model (HMM) [15], Support Vector Machine (SVM) [16], Maximum Entropy Model (MEM) [17] and Conditional Random Fields (CRF) [18], have been used to solve the information extraction problem. CRF has been proven to be the state-of-art model among these models. Taira and Soderland [19] first used MEM for the task of knowledge acquisition, parsing, semantic interpretation and evaluation of radiology reports; then, they moved to a vector space model to extract concepts about anatomy defined in the UMLS. A set of 2551 unique anatomical concepts was finally extracted from brain radiology reports, and an *F*-score of 87% was achieved [20]. Sibanda et al. [21] employed SVM trained with syntactic, contextual clues and ontological information from UMLS to recognize semantic categories in discharge summaries. They extracted eight types of semantic categories, and an *F*-score above 90% was achieved. There are also some methods of multiple classifier fusion for this task. For example, Wang and Patrick [22] combined MEM, SVM and CRF to recognize 10 types of clinical entities from 311 admission summaries, and an *F*-score of 83.3% which is 3.35% higher than the baseline stand-only CRF model, was obtained. Li et al. [23] compared CRF with SVM for disorder named entity recognition in clinical texts, and the experimental results showed that CRF obtained a higher score than did SVM.

All of the works described above are, however, not evaluated on the same data set, so it is difficult to compare them. The 2010 i2b2/VA challenge provides an opportunity for researchers to demonstrate their methods on a shared data set. Most of the submitted systems are based on machine learning methods. The best performance was achieved by a discriminative semi-Markov HMM that was trained by passive–aggressive (PA) online updates. The system obtained an *F*-score of 0.8523 [24]. Roberts and Harabagiu [25] proposed a flexible feature selection mechanism that makes it easy to find a near-optimal subset of features for a task in their system. For more details about this challenge, refer to [2]. There have also been some works based on the data set after this challenge. Xu et al. [26] developed a system that outperforms the best system in the challenge. Their main contribution to concept extraction was using two separate CRF models to handle medical concepts and non-medical concepts. Chen et al. [27] applied active learning to assertion classification, and their experiments showed that a comparable performance can be achieved with fewer annotated training instances. Abacha and Zweigenbaum [28] compared three methods of medical entity recognition: a semantic method that relies on domain knowledge, a method that first extracts noun phrases and then uses SVM to classify their entity types, and a method that uses CRF to identify entity boundaries and types simultaneously. Their work showed that the hybrid method that combined machine learning and domain knowledge yielded the best results.

Although statistical machine learning methods have obtained certain achievements, the lack of abundant annotated clinical documents and the diversity in clinical documents from multiple institutions present great challenges to researchers. One solution to this problem is to accomplish the task with fewer or even no training data. For example, Zhang and Elhadad [29] attempted an unsupervised method to extract named entities in both biological and clinical text without any rules or annotated data. The advantage of this method is that it is easy to use in different applications; however, it is not as competitive as supervised methods. Another solution is to achieve clinical concept extraction by increasing the training data. Torii et al. [30] found that the performance may be improved if more training data are available; however, they also found that the performance of a model trained on one institution's data degraded when data from another institution were tested. Their work inspires us to explore new machine learning methods to improve the performance of clinical concept extraction models with the help of training data from other sources with different distributions.

Please cite this article in press as: Lv X et al. Transfer learning based clinical concept extraction on data from multiple sources. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.05.006